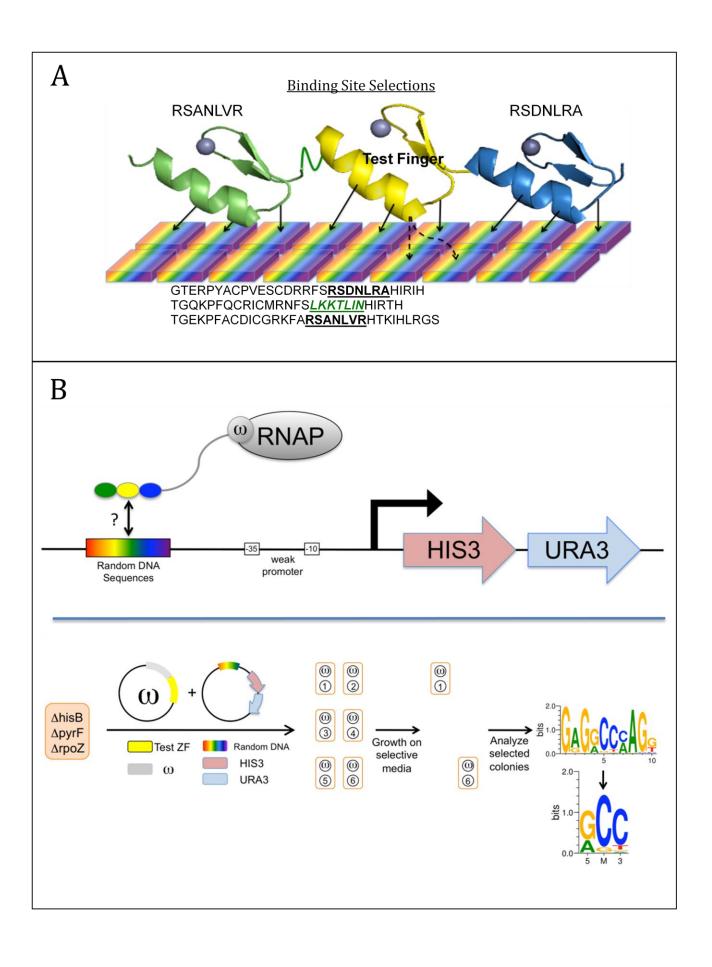
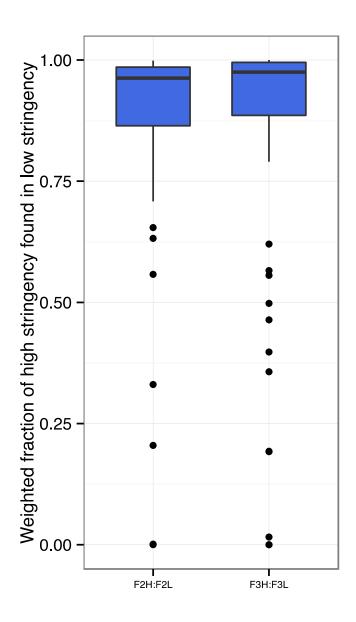
## **Supplemental Figures**

## Figure S1. Schematic of bacterial one-hybrid zinc finger binding site selections

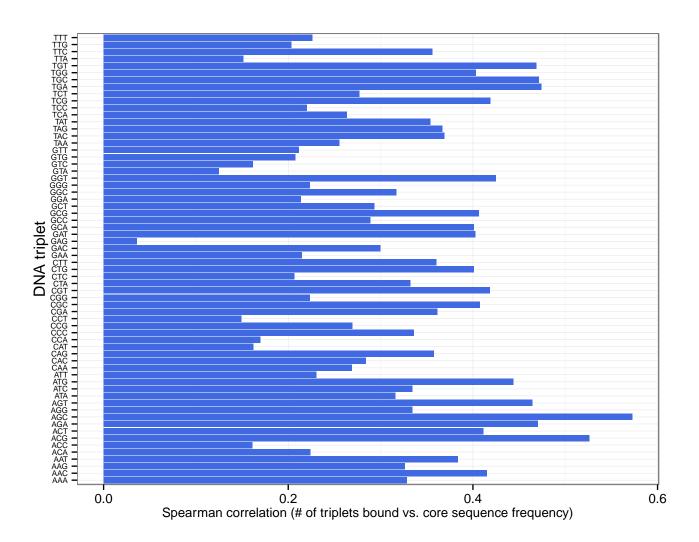
The DNA-binding specificities for candidate "test" zinc fingers were determined by selection of binding sites from a random DNA library as previously reported and described (see **Supplemental Methods 1d**). (a) The figure shows an example of a F2 candidate and the 9 random bases it may interact with (depicted as rainbow-colored bases). F3 candidates are characterized in the analogous way. The complete sequence of an example candidate is listed below the scheme, where the sequence of the test helix is shown in green. (b) Schematic of zinc finger binding site selection. (Top) Test zinc fingers are expressed as a 3-fingered protein-direct fusion to the omega subunit of RNA polymerase. Binding sites are selected from a 28 base pair region of random DNA sequences upstream of the promoter that drives the reporter genes, HIS3 and URA3. (Bottom) Two plasmids, the zinc finger expression vector and the random library reporter vector, are transformed into the bacterial strain. Double transformants are plated on selective media. DNA is recovered from bacteria that survive the selection and the binding site region of library reporter vector is recovered. After motif finding, selected binding sites are shown as a sequence logo. For simplicity, these have been trimmed to the 3 bases selected by the test finger for display in all other figures. (Next page.)



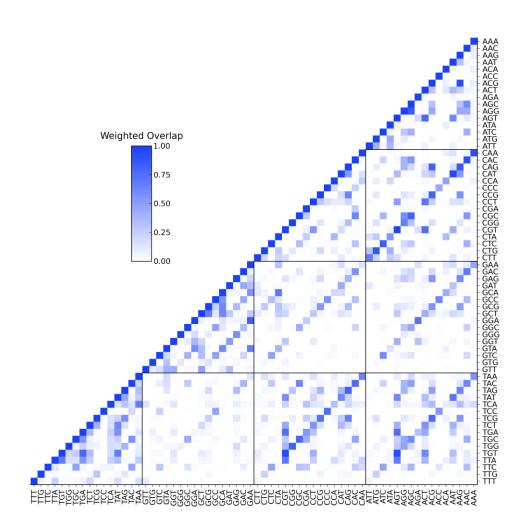
**Figure S2.** Agreement between high and low stringency protein selections. For each 3bp DNA target, we compute the weighted fraction of core sequences found in high stringency that are also found in low stringency in the F2 (left) and F3 (right) positions (see **Supplemental Methods 2b**). We visualize these weighted fractions across the 64 possible targets via boxplots, as in Figure 2b. For most targets, this weighted fraction is close to 1, with relatively small variance across the targets, but with a few notable outliers. Thus, for most targets in both F2 and F3, a large fraction of the frequency-weighted population of core sequences observed at high stringency is also observed at low stringency. Overall, the overlap between the independent high and low stringency selections shows the high quality and reproducibility of the data.



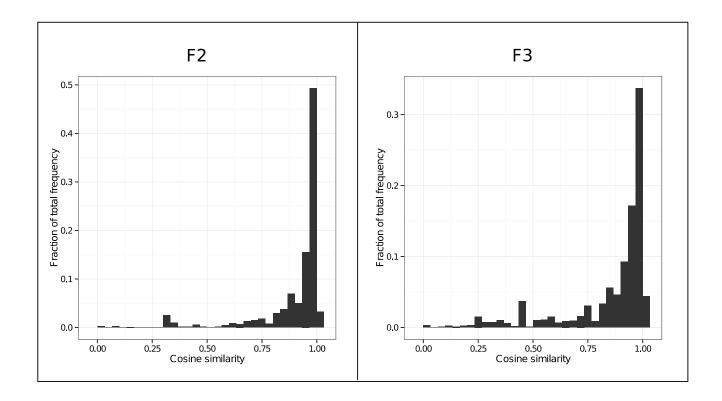
**Figure S3.** The number of triplets bound by a core sequence is correlated with its per-DNA-target frequency of binding. For each 3bp DNA target, we give the Spearman rank correlation coefficient between the frequency with which each core sequence was observed in the combined F2+F3 protein selections with the number of DNA targets bound by that core sequence across the dataset. For all targets, there is positive correlation between number of targets bound by a core sequence and its frequency of binding within an individual target. *P*-values obtained via a permutation test show all but three of these correlations to be statistically significant (FDR < 0.05 using the Benjamini-Hochberg procedure).



**Figure S4. Similar 3bp DNA targets tend to be bound by similar core sequences.** We show the frequency weighted overlap of the core sequences binding each pair of targets (see **Supplemental Methods 2b**) in the F3 protein selections as a heat map. The darkest shade of blue represents a score of 1 (complete overlap) and white represents a score of 0 (no overlap). As is apparent by the various patterns in the heat map (diagonal stripes and triangles along the upper diagonal), the target pairs showing highest degree of similarity tend to differ in only one base position.



**Figure S5.** Core sequences recovered from both high and low stringency protein selections tend to have similar binding profiles. For each of the core sequences that were selected at both high and low stringency (1,549 for F2 and 1,482 for F3), we compared their independently inferred binding profiles (computed as described in **Methods**) using the cosine similarity measure and plotted the total frequency of core sequences that fell into discrete similarity bins (**Supplemental Methods 2b**). The majority of the density for this distribution lies in the high-cosine similarity region. Thus, the independently performed high and low stringency selections show good agreement with respect to the binding profiles computed for core sequences.



**Figure S6. DNA-binding specificities inferred via lookup based upon protein selections show excellent correspondence with experimentally determined DNA-binding specificities.** For each of 166 helices from which DNA-binding specificities were successfully obtained in the F2 positional context, we give a paired entry with an experimentally determined logo (top) and a logo computationally inferred from a helix's binding profile in the F2 selections and the lookup procedure, as described in **Methods** (bottom). If the core sequence for a protein is not present in any of the F2 selections, there is no predicted specificity and this is indicated by the absence of a logo. Approximately 83% of the 498 (=166 x 3) paired nucleotide frequency vectors (experimental vs. inferred) are in good agreement (Pearson correlation coefficient >= 0.5) and >67% of the logos show agreement in all three base positions. (Next 3 pages.)

AASTLWH	ANRNLKE	AQSNLRA	AQSNLWR	CKPNLNR	CNRNLTE	CNRVLQD
E. CAA	E . GC	#1.0 CAT	£10 <b>0</b>	a o GAC	# . CCC	£ 10
	21.0	20		20 20 20 20 20 30 30 30	£1.0	3 · · · · · · · · · · · · · · · · · · ·
DQSALLG	DQSNLRA	DQSNLTR	DQSNLYH	DRSNLIK	EATSLRN	ELTNLKH
20 t.0 C V	20 1.0 C	20 1.0 0.0	20 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0	0.0 CA	± 1.0 A C	E AC
25 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1	25 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	# 1.0 GA	2.0 2.0 2.0 3.0 4.0 5.0 6.0 6.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7	2.0 2.0 2.0 3.0 4.0 5.0 6.0 5.0 6.0 6.0 6.0 6.0 6.0 6.0 6.0 6.0 6.0 6	± 1.0	£ C.
FASALYN	FASNLTD	FASNLTR	FASTLYN	FKHHLTR	FLCNLSR	FQSALTR
20 10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	E C C C C C C C C C C C C C C C C C C C	5 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0	20 10 A L	20 1.0 0.0 5 M 3	島 to 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	20 to 1 to
ACC NO.	E CO	20 GIA	B. C.	20 10 10 10 10 10 10 10 10 10 10 10 10 10	# 1.0 <b>CA</b>	20.00
FQSGLIS	FQSNLIA	FQSNLIR	FQSNLTR	FQSSLIY	FQSTLIQ	FQSTLSN
20.0	E LEG	#1.0 <b>\</b>	2 10 <b>1</b>	2 1.0 6.0 6.0 6.0	A TO ACT	210
23.0	210	#1.0 GIA	ETA	20 20 20 20 20 20 20 20 20 20 20 20 20 2	20 20 20 20 20 20 20 20 20 20 20 20 20 2	£1.0
HASGLCR	HASNLTR	HASSLCR	HASSLYN	HASTLTR	HKCTLTD	HKRNLTD
2 12 VIII	S C C	0.0 5 M 3	25 1.0 0.0 5 M 3	E S M 3	E LO CC	Z to CGG
25 25 1.0 0.0 5 5 M 5	SAC.	20 20 10 10 10 10 10 10 10 10 10 10 10 10 10		Z L CCC	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 25 to 20 00 2 M 3
HPSTLRY	HPSTLYY	HQANLIH	HQSALYN	HQSGLLG	HQSNLIN	HQSSLCR
	ACC ACC	2 1.0 0.0 0.0 0.0 0.0 0.0	A LOCAL COMPANY OF THE PARTY OF	A A	# ACC	8 to 10 to 1
20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 20 20 20 30 30 30 30 30 30 30 30 30 30 30 30 30	ACG	型 1.0 2.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3	g AGG	2
HQSTLTN	HRSGLYN	HRSNLLG	LKGNLTK	LKKTLIN	LKKTLTD	LNAHLTR
2 10 ACT	2 10 A	±1.0 VAC	2 10 S N S	E 10 GCC	±10 CC	£ 10
25 AC	20 10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	# 1.0 0.0 5 M 3	2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	20 20 00 00 00 00 00	# 10 <b>1</b>	
	PASALYN  FASALYN  FASALYN  FASALYN  FASALYN  FOSGLIS  FOS	POSALLG DOSNLRA   FASALYN FASNLTD   FOSGLIS FOSNLIA   FOSGLIS FOSN	DQSALLG DQSNLRA DQSNLTR  ### Company of the company	DOSALLG DOSNLRA DOSNLTR DOSNLYH  FASALYN FASNLTD FASNLTR FASTLYN  FOSGLIS FOSNLIA FOSNLIR FOSNLIR  HASGLCR HASNLTR HASSLCR HASSLYN  FOSGLIS FOSNLIA FOSNLIR FOSNLIR  HOSALTR HASSLYN  HOSALTR HAS	DQSALLG DQSNLRA DQSNLTR DQSNLYH DRSNLIK  FASALYN FASNLTD FASNLTR FASTLYN FKHHLTR  FQSGLIS FQSNLIA FQSNLIR FQSNLTR FQSSLIY  HASGLCR HASNLTR HASSLCR HASSLYN HASTLTR  HPSTLRY HPSTLYY HQANLIH HQSALYN HQSGLLG  HQSTLTN HRSGLYN HRSNLLG LKGNLTK LKKTLIN	DOSALLG DOSNLRA DOSNLTR DOSNLYH DRSNLIK EATSLRN  FASALYN FASNLTD FASNLTR FASTLYN FKHILTR FLCNLSR  FOSGLIS FOSNLIA FOSNLIR FOSNLTR FOSSLIY FOSTLIQ  FASALYN HASSLCR HASSLYN HASTLTR HKCTLTD  HASGLCR HASNLTR HASSLCR HASSLYN HASTLTR HKCTLTD  HOSTLTN HRSGLYN HOANLIH HOSALYN HOSGLLG HOSNLIN  FOSTLIN HRSGLYN HRSNLLG LKGNLTK LKKTLIN LKKTLTD

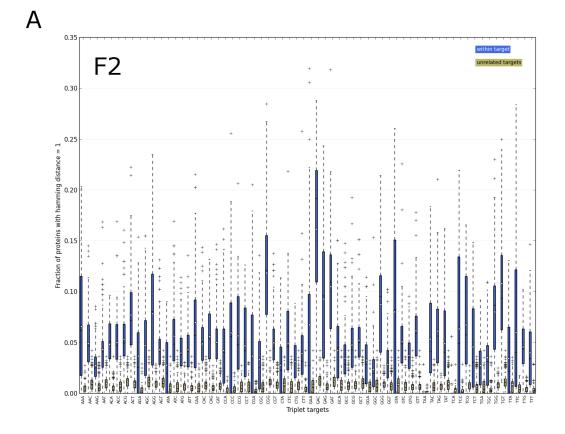
LNDHLQN	MKHHLTD	MKHHLTN	MKHHLTY	MKQHLTN	MKQHLTY	NPGTLWH	NPGTLYY
2 10 ACC	2 1.0 CU	E 10 C C C C C C C C C C C C C C C C C C	#1.0 CA	# AC	£10	#1.0 ACA	£ 10-
2010		E L	20 10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	£ 1.0	20 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	g 1.0 c 1	20 C
NSGALWN	QAAGLYN	QDCNLIR	QKAHLTM	QKAHLTR	QKAHLTT	QKGHLST	QKGILTN
20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 1.0 C	2012	200	± 1.2 CC	200	# 1.0 CA	Ž 10 CAC
20 22 to	20 10	20	20	20	20	± 1.0	20 20 20 20 20 20 20 20 20 20 20 20 20 2
QKGNLER	QKGNLKT	QKGNLRL	QKGSLTR	QKINLIN	QKVNLIR	QKVTLTH	QLATLSN
£ 10 10 10 10 10 10 10 10 10 10 10 10 10	20 A	20 10 00 00 00 00 00 00 00 00 00 00 00 00	20 GCA	£ 1.0	20 CAA	£ 1.0	B 12 CA
B. C.	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20	是1.0	£ 1.2		盘1.0	20 20 20 20 20 20 20 20 20 20 20 20 20 2
QLATLSY	QLATLTR	QLSTLNY	QSGNLRA	QSGTLTM	RKAGLYN	RKANLTM	RKASLVD
£ 1.0	2 10	2 1.0 A	21.0 CAA	210 AT	21.0 G G G G G G G G G G G G G G G G G G G	22 1.0 0.0 5 M 3	2 10 S M 3
£ . C.	23 10 10 10 10 10 10 10 10 10 10 10 10 10	Z10 CA	# 10-CA	20 20 20 3 M S	#10-0-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-	#10-144	E 10 C
RKATLTD	RKDHLVN	RKDSLVM	RKDTLAD	RKDTLTR	RKDYLTN	RKDYLVM	RKEDLKD
20 10 00 5 M 3	25 1.0 0.0 0.0 0.0 0.0 0.0	21.0	2 1.0 0.0 S M 3	20 CC S M 3	2 1.0 A	25 1.0 0.0 5 M 3	± C =
ž. C	22 12	20 20 20 20 20 20 20 20 20 20 20 20 20 2	20	± 1.2	20	20	
RKGDLKM	RKGDLTR	RKGNLKS	RKQTLTE	RKWTLIN	RKWTLKL	RKWTLVL	RKYHLVC
E VG	E CVA	20 S M 3	4 1.0 G	20 S M 3	型 1.0- 5 M 3	25 1.0 0.0 5 M 3	2 10 GG
RNAHLTD	20 10 10 10 10 10 10 10 10 10 10 10 10 10	200	20 CC	£ 1.0	型 1.0	至1.0 5 M 3	20 20 20 20 20 20 20 20 20 20 20 20 20 2
RNAHLTD	RNAHLTN	RNAHLVM	RNEHLTR	RNEHLVL		RRDTLRD	RRGDLKT
£ 1.6	ž 10 <b>Q</b>	£ 10 TG	£ 1.0 CC	±1.0 <b>TG</b>	RNGNLTR	£ 1.0	2 10 CVS
20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	23 GU	20 20 0.0 5 M 3	20 20 3 1.0 5 M 3	22 1.0 0.0 5 M S	\$1.0 G	# 1.0 0.0 5 M 3	20 A 3

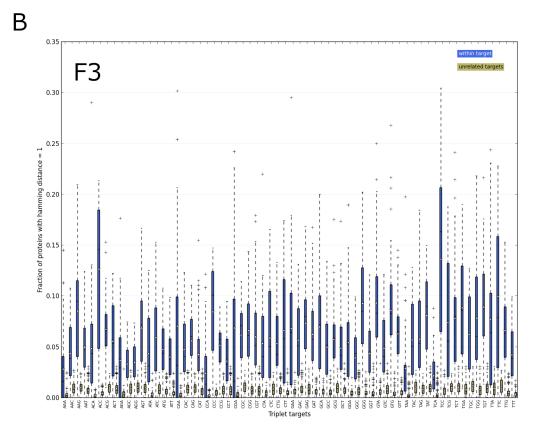
RRQHLSD	RSDNLKD	RVYELNR	SAGSLYN	SNRSLID	SQSGLCH	SQSGLYN	SQSNLLG
£10 - G	25 1.0 0.0 5 M S	21.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0	2 10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	#1.0 Tel	2 15 A N	#1.0 AT	End G
2 10 QG	B1.0	210 A	210	41.0 GG	20 10 10 10 10 10 10 10 10 10 10 10 10 10	210 A	20 00 00 00 00 00 00 00 00 00 00 00 00 0
SQSTLWH	SRQGLQI	SYNGLKG	TKFILDR	TKFVLTN	TKGNLKA	TKQHLTY	TKQILDN
EAC S M S	20 t.0	23 t.0. 6.0. 5 M 3	B 10 A 5	20 1.0. 0.0. 5 M 3	20 1.0 0.0 5 M 5	20 1.0 E I	20 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
20 20 20 5 M 3	20 20 20 20 20 20 20 20 20 20 20 20 20 2	20 CC	E L	# 1.0 0.0 5 M 3	22 A A A A A A A A A A A A A A A A A A		
TKQNLQS	TKQNLTH	TKQNLTN	TKQVLTE	TKRILSD	TKWNLRS	TKWNLTL	TKWTLDC
20 1.0 CO	20 S M 3	20 C C C C C C C C C C C C C C C C C C C	210	# 1.0 <b>GG</b>	2 1.5 CAG	型1.0 C	
20 C	20 20 20 20 30 40 50 40 50 50 50 50 50 50 50 50 50 50 50 50 50	20 10 C	B 10 E	20 20 20 20 20 20 20 20 20 20 20 20 20 2	20 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.	型 1.0 c	25 N 3
TKYDLTR	TKYILDN	TQSGLCQ	TRASLTP	TRRDLTR	TSSGLWS	VKRYLQA	WAHNLRD
2 1.0 CCC 0.0 N 3	# 1.0 AAG	2 1.0 A A	2 10 d d d d	2 1.0 CCC 6 M 3	2 15 15 15 15 15 15 15 15 15 15 15 15 15	#1.0 ACU	210
	20 20 30 30 30 30 30 30 30 30 30 30 30 30 30	20 20 20 310 40 40 40 40 40 40 40 40 40 40 40 40 40	20 00 00 00 00 00 00 00 00 00 00 00 00 0	£10000	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 20 20 20 20 20 20 20 20 20 20 20 20 2
WATSLTR	WHSNLID	WHSNLIN	WHSNLIR	WKGVLTR	WKLCLDL	WKRCLDI	WKSCLVL
20 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.	EGT	ä LAGC	310 <b>Q</b> C	10 GC	20 1 E	± 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0	20 1.0 L
20 CC	20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		2 L	20 10 CC		20 20 20 20 20 20 20 20 20 20 20 20 20 2	23 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2 1.2
WKSSLKD	WKTSLTD	WLSSLAY	WLSSLTN	WQSALIQ	WQSALTR	WQSSLIH	WQSSLIN
2 1.0 6 0 5 M 5	25 1.0 CA	20 1.0 CLD 5 M 3	S 10 S M 3	1.0 AT	10 10 1 E	25 1.0 A S	20 1.0 A S M S
200 CC	B 1.0	S. GA	20 10 10 10 10 10 10 10 10 10 10 10 10 10	\$ 1.0 A	20 10 1C	\$1.0	20 20 20 20 20 20 20 20 20 20 20 20 20 2
WQSSLTR	WRFSLDV	YKCNLLS	YKSVLTN	YQPNLIN	YRWILRD		
2 10 E	22 1.0 G	23 1.0 C	2 10 GI	22 1.0 0.0 5 M 3	2 10 0 0 M 3		
20 C C C C C C C C C C C C C C C C C C C	20 20 30 5 M 3	20 LAI	20 10 A 3		20 10 10 10 10 10 10 10 10 10 10 10 10 10		

**Figure S7.** A diverse set of helices are found to specify each 3bp target. For each 3bp target (organized as in Figure 4 in the main text), we visualize via sequence logos all the helices observed in our selections that are computationally predicted to specify it via lookup of binding profiles (described in **Methods**). Helices from F2 selections are shown on the left, and from F3 selections are shown on the right. Sequence logos show all 6 variable amino acid positions and the fixed Leucine in position 4 of the helix. Within each logo, the frequency with which each helix is observed across all selections (i.e., the sum of the per-target frequencies computed in **Supplemental Methods 2a**) is used to weight that helix's contribution to the logo.

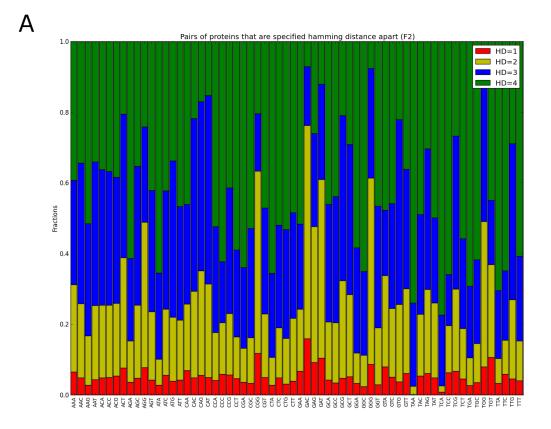
## middle base Т С Α G 230 KS T 230 KS T 2500 KS \$30 \$20 10 200 S \$30 VEWT BD \$30 VES B\$ Т С Т Α G Т С С #3.0 #3.0 #3.0 PUT ES #300 KU TS #300 EU S #20 OKS LD #20 SYNC EG Α base $\omega$ G 2 300 SS N 300 ST R8 Т С Α Α G Т С G 30 0 G G B 20 0 G G Α 300 NDT N 200 RS 100 RS 300 K Y B 300 E S 1 G

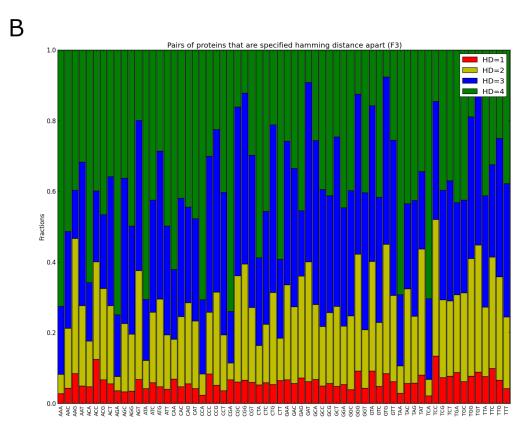
**Figure S8. Similar core sequences are found specifying each 3bp target.** For each 3bp target, for each core sequence, C, assigned to it via lookup using either F2 (top, (a)) or F3 (bottom, (b)), we computed the fraction of other core sequences assigned to that target that were identical to C in 3 of the 4 core positions of the recognition helix, and displayed this as a boxplot (shown in blue). For each target, T, for each core sequence assigned to it, we also computed this fraction across targets that do not share a common base with T in any of the three positions, and displayed this as a boxplot (shown in brown). A lower fraction of core sequences identical in 3 of 4 positions were found when looking across unrelated targets (124 out of 128 F2 and F3 targets were statistically significant at FDR < 0.05, with p-values obtained via the Mann-Whitney U-test and FDR correction via the Benjamini-Hochberg procedure). (Next page.)



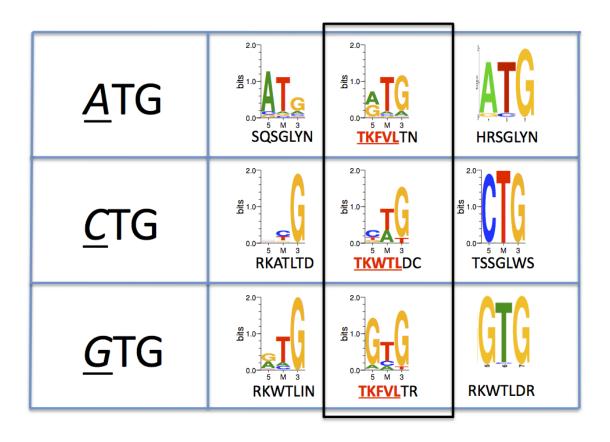


**Figure S9.** Both similar and dissimilar core sequences are found to specify each 3bp target. For each 3bp target, we computed the Hamming distance between all pairs of core sequences assigned to it via lookup using either F2 (top, (a)) or F3 (bottom, (b)) protein selections. For each target in both positions, core sequences that are inferred to specify it show a range of similarities, from pairs differing in only 1 of 4 amino acids in the core positions of the C2H2-ZF domain to pairs differing in all 4 amino acids in the core positions. (Next page.)

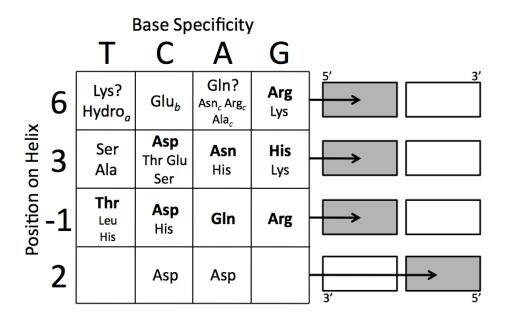




**Figure S10. C2H2-ZF domains with similar and dissimilar sequences selected to bind related targets.** Zinc fingers selected to bind common 3bp targets enrich for multiple, dissimilar solutions. Representative candidates tested to bind the related <u>A</u>TG, <u>C</u>TG, and <u>G</u>TG targets are shown. C2H2-ZF domains with dissimilar sequences are able to bind each target as demonstrated by the determined specificities (test helix sequences are noted below each logo, N to C). In fact, 3 or 4 of the core sequence residues may be different, yet the helices still bind a common target such as the ATG and CTG examples shown, though the information content provided may also vary as a result. On the contrary, similar domains can bind different but similar targets where the differences are dictated by a difference in a single contact from the canonical binding model. For example, helices of the sequence trend **TK-aromatic-(T/V)L-**<u>XX</u> are tested in each case (boxed column). All of these specify 5' n-TG 3' as expected. The 5' "n" base is dictated by the <u>XX</u> amino acids that distinguish these domains.



**Figure S11.** A recreation of the zinc finger code offered in the paper by Wolfe and Pabo noted in the figure. Within each box, the amino acid, or type of amino acid, that is predicted to specify a given base (for a given position of the helix, according to the canonical model) is noted. Core positions of the recognition helix are noted to the left of their specified rows. The predicted base preference is denoted by column, labeled above. The position within the 3-4bp target which is contacted by the specified helical position is shown to the right with an arrow. Additions to the code coming from various literature sources are also noted.



## Table modified from:

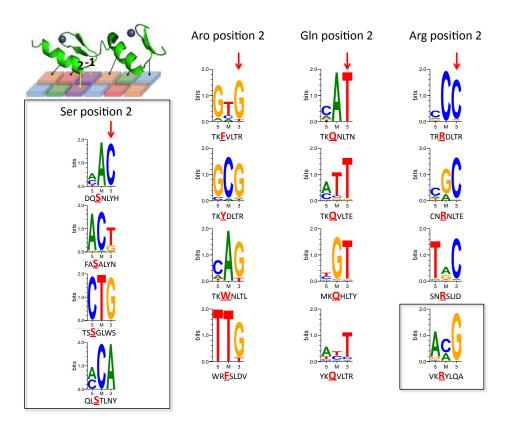
Wolfe, S.A., Nekludova, L. and Pabo, C.O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct*, **29**, 183-212. Additions:

a Maeder, M.L., Thibodeau-Beganny, S., Osiak, A., Wright, D.A. et al. (2008) Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. Mol Cell, 31, 294-301.

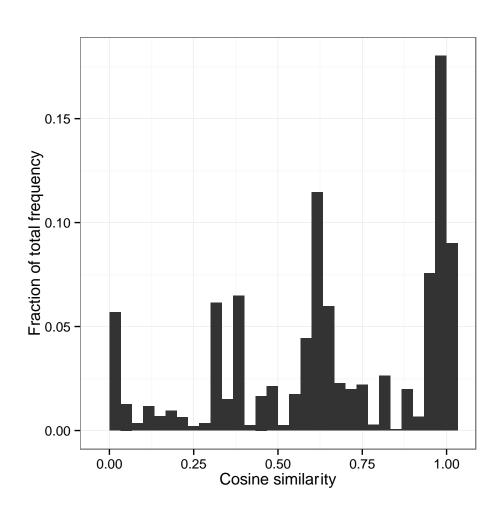
b Dreier, B., Fuller, R.P., Segal, D.J., Lund, C.V., Blancafort, P., Huber, A., Koksch, B. and Barbas, C.F., 3rd. (2005) Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors. *J Biol Chem*, **280**, 35588-35597.

c Dreier, B., Beerli, R.R., Segal, D.J., Flippin, J.D. and Barbas, C.F., 3rd. (2001) Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. *J Biol Chem*, **276**, 29466-29478.

Figure S12. Intra-3mer contact at *b3* made by position 2 of the recognition helix. Zinc finger selections reveal common trends between amino acids at position 2 of the helix and the specified 3' base, or as labeled in Figure 1a, "*b3*". (Right) For example, in the F2 protein selections, aromatics at position 2 are enriched for some targets with Guanine at b3. Glutamine is enriched for some targets with Thymine at b3 and Arginine is enriched for targets with Cytosine at b3. Subsequent characterization of DNA-binding specificity confirms these preferences. Representative examples are shown with the position 2 amino acid bold and red in the helical sequence listed below its binding logo. A red arrow points to the b3 position in each column. One exception to the Arg2 trend is shown at the bottom right (VKRYLQA) where the Arginine does not lead to b3 Cytosine specificity, which may be influenced by the aromatic at position 3. It should be noted that in all position 2 examples shown there are one or more helices with a Threonine at position -1 while the base preference trends with the position 2 amino acid base preference indicated. (Left, box) A contrary example, where a position 2 amino acid that does not trend with b3 specificity (Serine) is shown. Example Ser2 fingers can show preference for any b3 base. Together, these examples suggest that, positions -1 and 2 influence b3 specificity and a combination of the amino acids at -1 and 2 may contribute to b3 base preference.



**Figure S13**. Core sequences uncovered in both F2 and F3 selections often have dissimilar binding profiles. We examined the set of core sequences that were selected at both high and low stringency in both F2 and in F3 and that also showed good correspondence with respect to their binding profiles at high and low stringency for each position separately (cosine similarity >= 0.25 in the Figure S6 comparisons). For each core sequence in this set, we compared the binding profile inferred from the F2 protein selections against the binding profile inferred from the F3 protein selections via cosine similarity and plotted the total frequency of core sequences that fell into discrete similarity bins (**Supplemental Methods 2b**). In contrast to the comparisons of low and high stringency data (Figure S5), it is apparent that while much of the density of this histogram still lies in the high cosine similarity bins, a large proportion of it lies in the low to mid-range cosine similarity bins as well. Thus, while there is good correspondence between many binding profiles derived from the F2 vs. F3 contexts, there are also many instances in which the change of positional context leads to starkly different inferred DNA-binding specificities.



**Figure S14.** Computationally inferred logos show the least amount of agreement in base **position 1** when comparing the F2 vs. F3 context. For each core sequence found to show good agreement between the high and low stringency protein selections within both the F2 and F3 contexts (see Figure S15 caption), we predicted DNA-binding specificities via lookup of the binding profiles (described in **Methods**) using either the F2 or F3 protein selections. For each of these core sequences, the resulting base-frequency distributions for each of the base positions (1, 2, and 3) were compared (F2 vs. F3) using the Pearson correlation coefficient (PCC), and displayed via boxplots. The lowest agreement is observed at base position 1, and these agreements are significantly lower than those observed in base positions 2 and 3 (p < 0.03 and p < 0.0001, respectively, as computed by the Mann-Whitney U-test). This is notable as this position in the subsite for F3 corresponds to the 5' most base of the entire binding site, and in the subsite for F2 may participate in a cross-strand contact with F3.

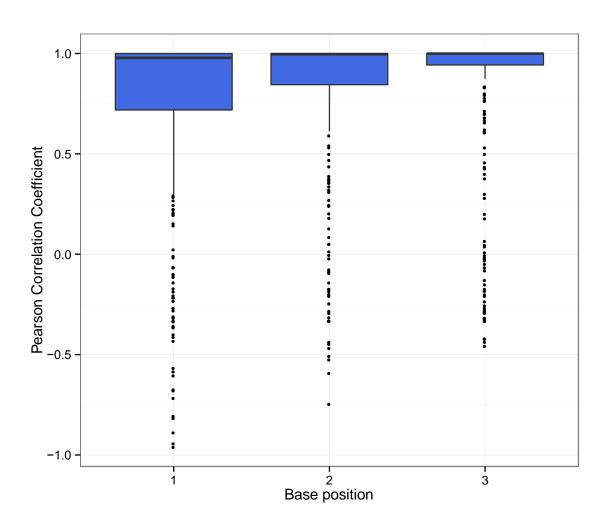


Figure S15. Several of the core sequences tested in both the F2 and F3 positions only exhibit **DNA-binding in one of the two positional contexts.** (A) Experimental binding site selections were performed in both the F2 and F3 positional context for 26 helices that were identical in the core positions of the recognition helix (-1, 2, 3, and 6). The derived logos show good agreement between the F2 and F3 context for 16 of these helices (top). However, 5 helices showed strong base preferences in the F2 context, but little or no functional DNA-interaction in F3 context (lower left; sequence marked with asterisk produced weak base preferences in the F3 context). Similarly, the remaining 5 helices show good base preferences in the F3 context, but give no evidence of functional interaction in the F2 context (lower right). (B) We produced computationally inferred logos via lookup (see **Methods**) for these same core sequences, based upon either the F2 or F3 protein selection data (F2 logos displayed above F3 logos). These logos show rough agreement between the F2 and F3 context for 13 out of the 16 of the core sequences that showed agreement experimentally (top). Additionally, differential DNA-binding behavior between the F2 and F3 positional contexts was successfully predicted (bottom left and right): in 9 out of 10 cases, no logo is predicted for one of the contexts due to insufficient information in the corresponding protein selection data. This reflects positional biases in protein selections that are analogous to those observed in binding site selections. (Next page.)

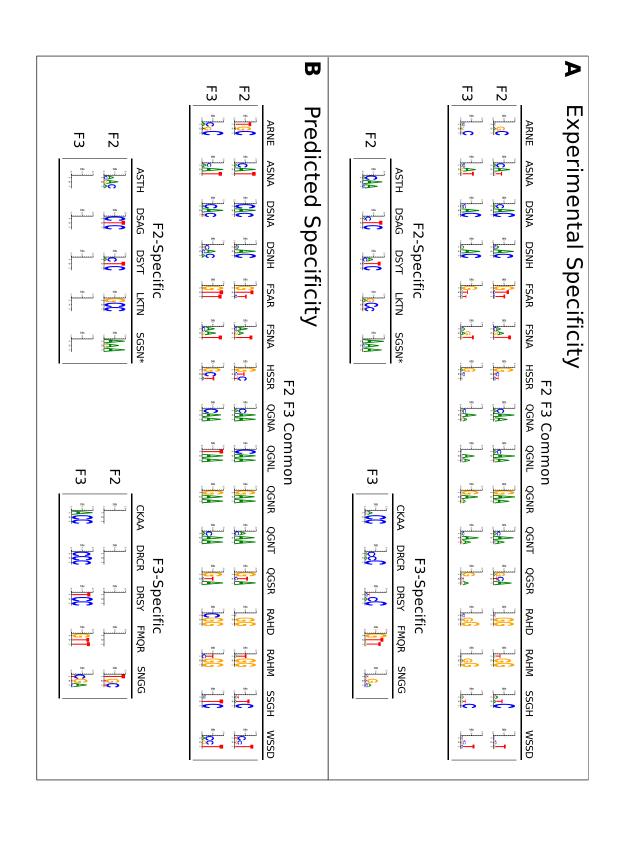


Figure S16. Selected C2H2-ZF assemblies that specify challenging targets.

Three-fingered arrays of C2H2-ZFs were selected from pool-derived libraries to bind targets that arrays constructed by modular assembly failed to recognize. (**Left**) Zinc finger numbers as specified by (Lam et al. 2011) are listed to the left of each row to provide a reference. (**First column**) For each target, the DNA-binding specificity of a selected three-fingered array was determined via B1H DNA-binding site selections and displayed as a sequence logo. Above the logo, the seven amino acids (-1,1,2,3,4,5, and 6) of each helix of the selected three-fingered array are listed N to C, for fingers F3, F2 and F1. Amino acids not coded for in VNS libraries are noted as bold and red. Below each logo, the desired target is provided in colored letters, 5' to 3'. A match between the preferred base of the sequence logo and the desired base in the target is indicated by a capital letter. (**Second column**) The activity of each selected zinc finger array, noted in the first column, was tested in yeast using an affinity-related GFP assay. The zinc finger arrays were challenged to bind the desired target (Correct) or a negative control consisting of an empty vector (Control). Fluorescence was normalized to a positive control and results are displayed as a percentage of this positive control activity.

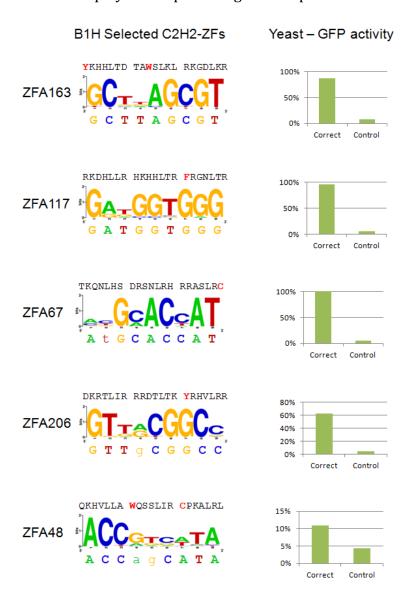


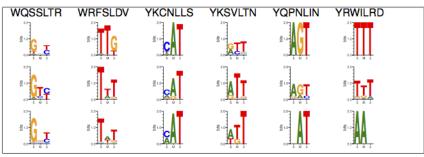
Figure S17. DNA-binding specificities predicted via nearest neighbor decomposition show agreement with experimentally determined DNA-binding specificities. (a) For each of 166 binding site selections performed in the F2 context, we give an experimentally determined logo (top), a logo computationally predicted by nearest neighbor decomposition using the F2 protein selections (middle; ignoring exact matches if present), and a logo computationally predicted by nearest neighbor decomposition using the F3 protein selections (bottom). (b) For each of the 69 binding site selections performed in the F3 context, we give an experimentally determined logo (top), a logo computationally predicted by nearest neighbor decomposition using the F3 protein selections (middle; ignoring exact matches if present), and a logo computationally predicted by nearest neighbor decomposition using the F2 protein selections (bottom). An empty logo corresponds to a case where nearest neighbor decomposition did not predict a specificity, due to no helices in our data set that had the required sequence similarity. (Next 8 pages.)

AAKTLWN	AASTLWH	ANRNLKE	AQSNLRA	AQSNLWR	CKPNLNR	CNRNLTE	CNRVLQD
2 to 10 to 1	20 10 CAA	Z to GC	21.0 CAT	210 <b>A</b>	E 10 GAC	20 cc	2 10 TC
2 1 0 1 C	20 20 20 20 20 20 20 20 20 20 20 20 20 2	2 . J.C.	2 10 CAT	Z. GAT	gAc	20 20 20 31.0 43.10	g g
E CG	20 <b>C</b>	20 C	2 10 A	200	E 10 CAC	E 10 CCC	ŽĮ V <mark>AG</mark>
DKSYLYT	DQSALLG	DQSNLRA	DQSNLTR	DQSNLYH	DRSNLIK	EATSLRN	ELTNLKH
2 C C C C C C C C C C C C C C C C C C C	20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Z to C	2 1.0 5 M 3	2 10 A	2 C C	22 1.0 A C	ž. AC
20 L	±1.0 CT	£1.0	E 10 GAÇ	20 10 E	# 1.0 0.0 0.0 0.0 0.0	± 1.0	20 CA
20 20 20 6 M 3	20 20 20 20 20 20 20 20 20 20 20 20 20 2	2.0 <b></b>	#10 CAC	B	25 25 25 35 35 35	#1.0 E 1.0 S M 3	20 20 20 5 M 5
FAASLYN	FASALYN	FASNLTD	FASNLTR	FASTLYN	FKHHLTR	FLCNLSR	FQSALTR
20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 1.0 Li	g cg	#10 <b>A</b>	25 to 15 to	20 5 M 3	£ 1.0 5 M 3	20 to 10 to
20 LC T	20 A L L	g col	£ 1.0	2012	20 GG	£ 1.0	20 E M 3
20 A A A A A A A A A A A A A A A A A A A	20 20 20 1.0 5 M 3	20 <b>G</b>	Z 1.0 5 M 3	20 00 00 00 00 00 00 00 00 00 00 00 00 0	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 A A A A A A A A A A A A A A A A A A A	25 A
FQSGLIQ	FQSGLIS	FQSNLIA	FQSNLIR	FQSNLTR	FQSSLIY	FQSTLIQ	FQSTLSN
2 10 A 1	20.0	a eg	2 10 A	Zas AI	2 1.0 ACT	ACT OF M 3	E to
2 10	2 1.0 A	a eg	2 10 GA	2	2 1.0 ACT	ACT	B. ACT
200	B: 0 G	20 A	£ 10	20	g 1.0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	E 10 AC	B1.0
FQSTLSR	HASGLCR	HASNLTR	HASSLCR	HASSLYN	HASTLTR	HKCTLTD	HKRNLTD
1 1 -	207 207 3	E N	10 C	2 10 10 10 10 10 10 10 10 10 10 10 10 10	# 1.0 GC	200 S M S	E COC
and the GO	20 1.0 TC	22 CAC	20 10 10 10 10 10 10 10 10 10 10 10 10 10	B 10 AC	25 1.0 GCC	20 20 20 20 20 20 20 20 20 20 20 20 20 2	
# 10 C	20 20 20 20 20 20 20 20 20 20 20 20 20 2	20 00 00 00 00 00 00 00 00 00 00 00 00 0	2 10 CT	G C C C C C C C C C C C C C C C C C C C	20 CL	20 5 W 3	20 CGC

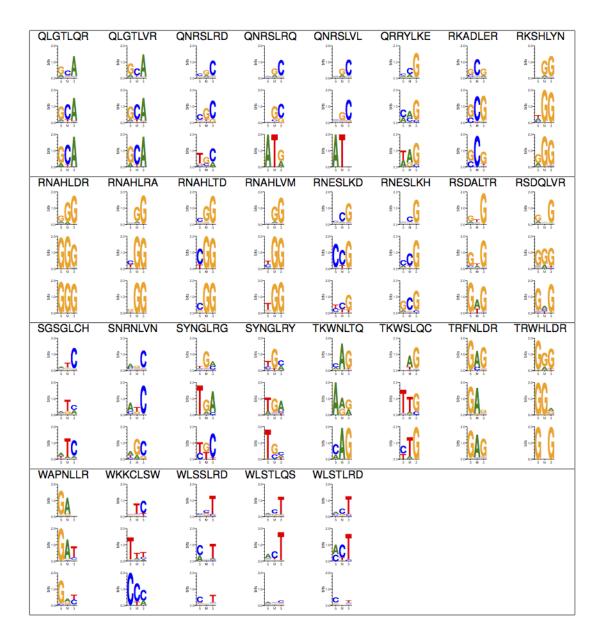
HPSTLGN	HPSTLRY	HPSTLYY	HQANLIH	HQSALYN	HQSGLLG	HQSNLIN	HQSSLCR
ACT	2 12 A S	2 1.0 ACC	210 A	Bro-ACG	2 10 A	AAC CGC	10 CT
2 · · · · · · · · · · · · · · · · · · ·	23 12 10 CC	20 CC	200	20 20 20 20 20 20 20 20 20 20 20 20 20 2	20 20 20 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40	AGC 0.0	20 10 10 10 10 10 10 10 10 10 10 10 10 10
2.5 ACT	2 2 2		210	JJ\$	£ 1.0	£10	£1.0
HQSSLCR	HQSTLTN	HRSGLYN	HRSNLLG	LKGNLTK	LKKTLIN	LKKTLTD	LNAHLTR
± 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	E S	20 A A A	ž ie C	25 to 5 M S	E LO GCC	210 CCC	22 1.0 0.0 5 M 3
22 1.0 1 C C C C C C C C C C C C C C C C C C	22 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	20 20 10 00 00 00 00 00	E 10 CAC	20 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.	20 20 20 20 30 5 M 3	20 10 C	20 20 20 5 M 3
2 1.0 L	a ACT	2 1.0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	ELCAC	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 10 10 10 10 10 10 10 10 10 10 10 10 10	25 10 COT
LNDHLQN	MKHHLTD	MKHHLTN 2003	MKHHLTY	MKQHLTN 2°3	MKQHLTY 2073	NPGTLWH	NPGTLYY
20 10 ACC	B 10 CV	E 10 SCG	E	# AC	£ 1.0 5 1	\$1.0 <b>ACA</b>	25 1.0 A A A
B LO GO	B sa CUT	20 20 3	201.0	20 GG-	20 20 20 3	E LO CA	20 CA
26 20 00 5 M 5	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 10 10 10 10 10 10 10 10 10 10 10 10 10	B. C. G.	20 20 20 30 5 M 3	型: 1.0 5 M 3	型 1.0 mm 3	型1.0 S M 3
NSGALWN	QAAGLYN	QDCNLIR	QKAHLTM	QKAHLTR	QKAHLTT	QKGHLST	QKGILTN 2071
210	2 10 A S	2 10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	E 10 GA	E to GA	2 10 A	E LO AÇA	E to CAC
200	20 12 A A S	2 1 2 1 A	2.0	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 ACA	20 AGA	20 AAA
2.0 2.0 3.0 5.0 3.0	20 AIA	20	2 to the second	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 A	20 A	20 20 20 3
QKGNLER	QKGNLKT	QKGNLRL	QKGSLTR	QKINLIN	QKVNLIR	QKVTLTH	QLATLSN
20 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	£ 12	20 20 00 00 00 00	26 CC	25 to 100	± 1.0 GAA	± 1.0 CA	22 1.0 0.0 5 M 3
and the second s	12 C A A A A A A A A A A A A A A A A A A	10 CO	20 00 00 00 00 00 00 00 00 00 00 00 00 0	25 AA	20 00 20 30 30 30 30 30 30 30 30 30 30 30 30 30	#1000000000000000000000000000000000000	10 ACA 00 5 M 3
2 to 1 to	20 10 20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 20 20 30 30 30 30 30 30 30 30 30 30 30 30 30	20 LA	20 A A A C S	20 20 2 10 2 10 2 10 10 10 10 10 10 10 10 10 10 10 10 10	20 GA	20 10 10 10 10 10 10 10 10 10 10 10 10 10

QLATLSY	QLATLTR	QLSTLNY	QSGNLRA	QSGTLTM	RKAGLYN	RKANLTM	RKASLVD
2 10 AAA	21.0 60 6 M 3	20 A A	2 1.0 CAA	2 1.0 0.0 5 M 3	Single G	A LO	2 1.0 0.0 8 M S
20 A	20 GCA	20 CA	20 E I I	ACA	26 N N N N N N N N N N N N N N N N N N N	£ 1.0	41.0 0.0 1 M S
E	Z 10 CA	20 20 20 3 M	20	20 CCA	E G G	20 A	20 LEV
RKATLTD	RKDHLVN	RKDSLVM	RKDTLAD	RKDTLTR	RKDYLTN	RKDYLVM	RKEDLKD
20 20 20 20 20 20 20 20 20 20 20 20 20 2	21.0 <b>CC</b>	22 1.0 CU	2 1.0 C	22 1.0 CCC	2 1.0 A	210 AG	±
20 20 00 20 00 00 00 00 00 00 00 00 00 0	E 10 GU	20 20 20 20 30 5 M 3	20 20 20 20 20 20 20 20 20 20 20 20 20 2	# 1.0 CG	S 10 AG	20 A	# 1.0 0.0 5 M S
20 20 20 20 20 20 20 20 20 20 20 20 20 2	20 20 20 20 20 20 20 20 20 20 20 20 20 2	20 CC	20 CG	20 GC	20 10 A 3	210 A	25 25 1.0 0.0 5 M S
RKGDLKM	RKGDLTR	RKGNLKS	RKQTLTE	RKWTLIN	RKWTLKL	RKWTLVL	RKYHLVC
\$ 1.0 0.0 2 M 3	Z. CVA	\$1.0 <b>C</b>	g <sub>1.0</sub> G	\$ 1.0 R	2 10 00 00 00 00 00 00 00 00 00 00 00 00	2 1.0 2.0 3 M 3	£ 1.0
20 10 10 10 10 10 10 10 10 10 10 10 10 10	#10 GG	型1.0 CAG	g	型 1.0	20 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.	型 1.0 量 1.0 5 N 3	20 20 20 20 20 3 3
20 20 20 3	# 10 <b>CC</b>	登1.0	20 G	至1.0	20 10 10 10 10 10 10 10 10 10 10 10 10 10	25 1.0 III	20 20 20 20 20 20 20 20 20 20 20 20 20 2
RNAHLTD	RNAHLTN	RNAHLVM	RNEHLTR	RNEHLVL	RNGNLTR	RRDTLRD	RRGDLKT
£10	10 00 00 00 00 00 00 00 00 00 00 00 00 0	25 1.0 TG	Sto CCC	10 10 10 10 10 10 10 10 10 10 10 10 10 1	210 10 10 10 10 10 10 10 10 10 10 10 10 1	M to Co	#1.0 <b>TUG</b>
20 CV	20 GU	20 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 10 10 10 10 10 10 10 10 10 10 10 10 10	A TE CAG	20 Sec. 3	20 AC
20 0 0 0	20 00 00 00 00 00 00 00 00 00 00 00 00 0	20 20 3 10 3 M S	20 GU	20 20 20 3 1.0 3 1.0 3 1.0	215 A 3	200 G	2010G
RRQHLSD	RSDNLKD	RVYELNR	SAGSLYN	SNRSLID	SQSGLCH	SQSGLYN	SQSNLLG
2 1.0 LGG	2 1.0 6.0 6 M 3	22 1.0 0.0 5 M 3	20 1.0 0.0 5 M 3	2 1.0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	20 10 10 10 10 10 10 10 10 10 10 10 10 10	210 AT	# 1.0 CA
220 CCT 000 S M S	Z O	20 20 00 5 M 3	20 A	20 20 20 3 1.0 3 M 3	20 TC	H 10 A 3	2.0 <b>VAC</b>
2.0 2.0 2.0 2.0 3.0 5.0 5.0 5.0	20 2 10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 20 00 5 M 3	20 210 00	20 20 20 20 20 20 20 20 20 20 20 20 20 2	20 AIA	20 CAL 5 M 5

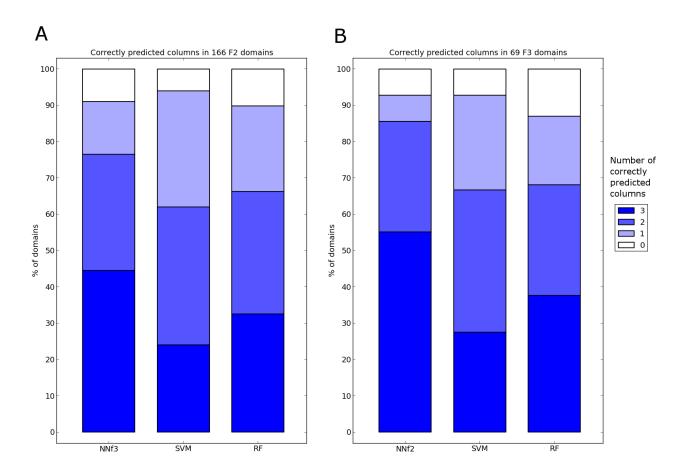
SQSTLWH	SRQGLQI	SYNGLKG	TKFILDR	TKFVLTN	TKGNLKA	TKQHLTY	TKQILDN
E CAC	\$1.0 CT	25 to 100	\$10 AG	E .	25 1.0-00 S M 3	End S	2 1.0 A T
Z. C	20 20 20 3 10	20 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	20 G	20 4 5	20 A	E . CGT	28 1.6 AG
20	£1.0	g <sub>10</sub>	£10 <b>G</b>	2:0	#1.0 CAI	g	25 21.5
TKQNLQS	TKQNLTH	TKQNLTN	TKQVLTE	TKRILSD	TKWNLRS	TKWNLTL	TKWTLDC
20 CAT	20 21.0 22.1.0 3.0.0 3.0.0	E 1.0	20 AT	E GGC	E 1.0 CAG	E so	2 to 1 to 1
£ 1.0	# 1.0 CAT	2.0 2.0 2.0 2.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3.0 3	# 1.0 m	£	± 10	B 10 C	2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1
20 A L	#1.0 C	20 A Z	#1.0- 0.0 5 M 3	2 2 3 4 5 M S	#1.0 0.0 5 M 3	E. TAG	E 10 S M S
TKYDLTR	TKYILDN	TQSGLCQ	TRASLTP	TRRDLTR	TSSGLWS	VKRYLQA	WAHNLRD
20 to CCC 5 w 3	AAG	20 A A	A 1.0-		# 1.0 0.0 5 M 3	ACU	20 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2010	#1.0 <b>AAG</b>	20 A	<b>20</b>	20 20 20 20 20 20 20 20 20 20 20 20 20 2	型1.0 20 3 M 3	20 20 20 20 20 20 20 20 20 20 20 20 20 2	EGT
Z. J. G.	20 20 10 5 M 3	20 20 20 20 20 20 20 20 20 20 20 20 20 2	g 10 G	B. C. GCC	20 20 20 20 20 20 20 20 20 20 20 20 20 2	AAG	R. L. C. C. S. M. S.
WATSLTR	WHSNLID	WHSNLIN	WHSNLIR	WKGVLTR	WKLCLDL	WKRCLDI	WKSCLVL
Į.	EST.O.	AGC	Sin JAc	S. GC	2 1.0 0.0 5 M 3	g vo	2 1.0 L
2.0 2.0 2.0 3.0 5.0 5.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7	20 COL	20 AG	#1.0 C	g of g	至1.0 至1.0	g	Act
200	E 10 CAT	20 A A	\$100 A	20 Jan 1	20 20 20 31.0 3 M 3	E 10	20 10 10 10 10 10 10 10 10 10 10 10 10 10
WKSSLKD	WKTSLTD	WLSSLAY	WLSSLTN	WQSALIQ	WQSALTR	WQSSLIH	WQSSLIN
20 to 00 to	100 CA 1	25 t.0	100 AC	Z C AI	# 1.0 T.	E C AC	2 1.0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
20 20 00 00 00 00 00 00	20 E I	20 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	20 A T	B	20 20 20 20 20 20 20 20 20 20 20 20 20 2	B	2 A T
2.0 2.0 T	£1.0 € T	20 20 20 20 20 20 40 40 40 40 40 40 40 40 40 40 40 40 40	£1.0	£ 1.0-	2.0 2.0 2.0	2.0 2.0 2.0	20 20 20 20
0.0 K	0.0 A L	0.0 <del>                                     </del>	.o. ♣ <u>∓</u>	00 15 0 5	ODNI IN VE	" <u>†</u>	0 E X X



ANRNLKE	AQSNLIA	AQSNLRT	AQSNLRT	ARRELRN	ASSNLRA	CPKALRA	DASNLIH
2 10 C	2 1.0 0.0	2 10 10 10 10 10 10 10 10 10 10 10 10 10	21.0 6.0 5 M 3	20 CC S M 3	21.0 2.0 3 M 3	210	Bro S
20 S	2 1.0 CA	2 1. O.	20 A 10 A 3	£10-	20 10 E	200	
21.0	20 000	E TO CAT	20 A	20 20 31.0- 0.0-5 M 3	20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	20 20 20 30 40 50 50	20 3
DKRWLIK	DPRCLRA	DPRCLVR	DPRSLRS	DPRSLVY	DPSNLRA	DQANLIH	DQANLIK
2 10 00 00 00 00 00 00 00 00 00 00 00 00	E 1.0 CCV	E 1.0 CC	10 CC	1.0 0.0 5 M 3	2 1.0 0.0 6 M 3	2 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	25 to 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
± 1.0	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 TCV	20 20 20 30 30 30 30	E LAC	E 10 AAC	22 1.0 A
20 1.0 20 5 M 3	20 210 00 5 M 3	20 G	20 20 00 5 M 3	20 20 00 5 M 3	20000	210	200
DQANLRA	DQANLTR	DQSSLIR	FHSNLCA	FQGSLQR	FQSALTR	FQSCLIR	FQSSLCR
2 10 AC	2 1.0 0.0 0.0 0.0 0.0	E C	A LO	200 J.T.	2 1. OTT	B 10 OST	B12 G T
20 CAC	20 10 10 10 10 10 10 10 10 10 10 10 10 10	25 12 12 12 12 12 12 12 12 12 12 12 12 12	型10 20 20 20 20 20 20 20 20 20 20 20 20 20	£10	20 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 10 10 1
2.0 2.0 2.0 2.0 2.0 3.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4	20 1.0 EAC	E AT	H 10 AAT	型1.0	20 10 10 10 10 10 10 10 10 10 10 10 10 10	2010	20 00 00 00 00 00 00 00 00 00 00 00 00 0
FSYHLRR	HAASLWN	HASSLWR	LKGNLKA <sup>20</sup> 7	LKMNLIR 2071	LLGNLQR	MKQHLTN 2°7	QKGHLAY
E C C C C C C C C C C C C C C C C C C C	ALCA ACA	2 10 C	10 C	61.0 GA	21.0 GA	g in G	2 1.0 <b>GA</b>
2.0 <b>C</b>			210 CAC	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 21 10 0.0 0.0 0.0 0.0 0.0	20 10 10 10 10 10 10 10 10 10 10 10 10 10	g CGA
2 1.0 GG	21. AC	S 18 S S S S S S S S S S S S S S S S S S		20 A S	A S	GGT.	EG
QKGHLRS	QKGHLTR	QKVHLQR	QLGNLRA 2°]	QLGNLRL 2071	QLGNLRT	QLGNLTR	QLGSLAR
20 20 00 6 M 3	1	1	E AA	#1.00 AA	2 10 AA		23 A
Harmonia Control of the Control of t	E COL	Harton Control of the	£10000	20 22 1.0 5 M 3	20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	£10	10 2 2 4 3 3 4 4 3 5 4 4 4 5 4 5 4 5 4 5 5 5 6 6 6 6 6 6 6
£ 1.0	20 10 10 10 10 10 10 10 10 10 10 10 10 10	20 20 20 20 20 20 20 20 20 20 20 20 20 2	20 60 50 310 60 50 50 50 50 50 50 50 50 50 50 50 50 50	25 1.0 0.0 5 M 3	20 20 20 20 20 20 20 20 20 20 20 20 20 2	200	20 10 10 10 10 10 10 10 10 10 10 10 10 10



**Figure S18.** Predictions based on nearest neighbor decomposition (NN) effectively extend the data to account for differences in positional context. The fraction of predictions with >= 3, 2, 1 or 0 columns correct (as judged by a PCC >= 0.5) is shown for the NN approach (described in **Methods**) and compared to two previous prediction approaches based on random forests (RF) or support-vector machines (SVM). (**A**) Results for the 166 C2H2-ZFs that were tested as F2 in our B1H DNA-binding site selections. NN predictions were computed based on the F3 protein selections. (**B**) Results for the 69 C2H2-ZFs that were tested as F3 in our B1H DNA-binding site selections. NN predictions were computed based on the F2 protein selections.



**Figure S19. Predictions based on nearest neighbor decomposition outperform the lookup approach that is based solely a single core sequence.** We compare the performance of our original lookup inference approach (based solely on the binding profile of the exact core sequence) with the performance of our nearest neighbor decomposition approach (both approaches described in **Methods**). (A, left) Fraction of correctly predicted per-nucleotide base preferences, as judged by a PCC >= 0.5, on the 166 helices tested in the F2 positional context based upon either F2 or F3 protein selections using either lookup or nearest neighbor decomposition (ignoring exact matches for predictions based on F2 protein selection data). (A, right) Fraction of these predicted 3bp binding specificities that have 0, 1, 2, or 3 base preferences correctly predicted. (B, left) Fraction of correctly predicted per-nucleotide base preferences, as judged by a PCC >= 0.5, on the 69 helices tested in the F3 position context based upon either F2 or F3 protein selections using either lookup or nearest neighbor decomposition (ignoring exact matches for predictions based on F3 protein selections data). (B, right) Fraction of these predicted 3bp binding specificities that have 0, 1, 2, or 3 base preferences correctly predicted.

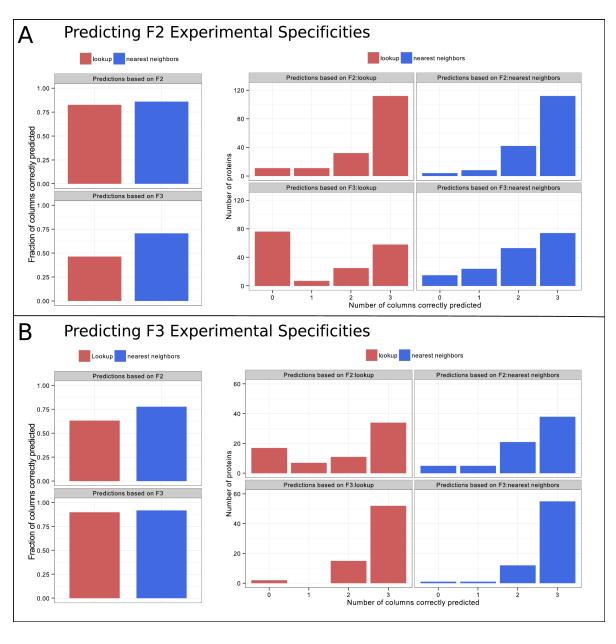


Figure S20. The nearest neighbor decomposition method (NN) performs comparably to other state-of-the-art C2H2-ZF binding site prediction methods (SVM and RF) for predicting binding sites of naturally occurring C2H2-ZF arrays. (A) The percent of proteins (y-axis) with statistically significant alignments between the predicted and experimental PWMs (Supplemental Methods 2e), shown across various p-value thresholds (x-axis). (B) The percent of proteins (y-axis) whose aligned predicted and experimental PWMs report at least a given percent of columns correctly predicted (x-axis), using an PCC threshold of 0.5 to determine whether an aligned column is correct. For both panels, NN is shown in green, RF in black and SVM in blue.

